A Scalable, Automated System for Photochemical Modeling on the Cloud

Matthew J. Alvarado, Erik J. Fanny, Ethan H. Fahy, Chantelle R. Lonsdale, and Elizabeth S. Bettencourt

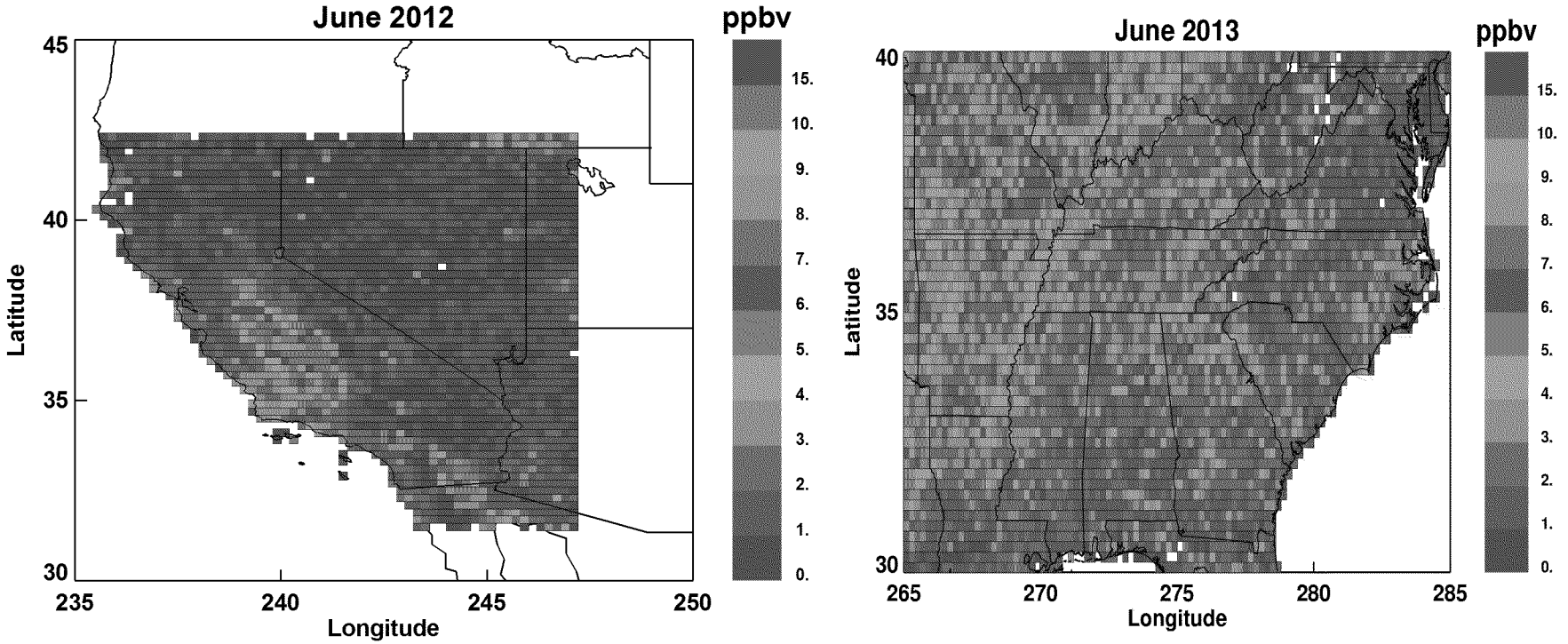
Atmospheric and Environmental Research (AER)
131 Hartwell Ave., Lexington, MA, 02421

A&WMA's 110th Annual Conference and Exhibition
Pittsburgh, Pennsylvania
June 5-8, 2017
Abstract #264612

aer
Atmospheric and Environmental Research

Our Scientific Goal: Evaluate NH_3 Emission Inventories Using Satellite Data

CrIS Monthly Average NH_3 retrievals



But going from idea to model output requires more fighting with software than actual science

For example, to run the bi-directional NH_3 flux components of CMAQ, simply:

1. Install and run FEST-C.
2. But to do that, you need to install the Spatial Allocator tools first.
3. And install VERDI.
4. And edit the "config.properties" file for your system.
5. And the "festc" file.
6. And the "festc_setup.csh" file.
7. And now modify your ".cshrc" file to source that file.
8. And download files on agricultural activity and soil properties (not included).
9. And adjust the format of the FEST-C output netCDF files so they will work in CMAQ.
10. And if you have to change machines, do it all over again.
11. Which you will need to, because it turns out this one doesn't have enough memory, or is too slow, or has an unsupported OS.

Our problems with the old approach

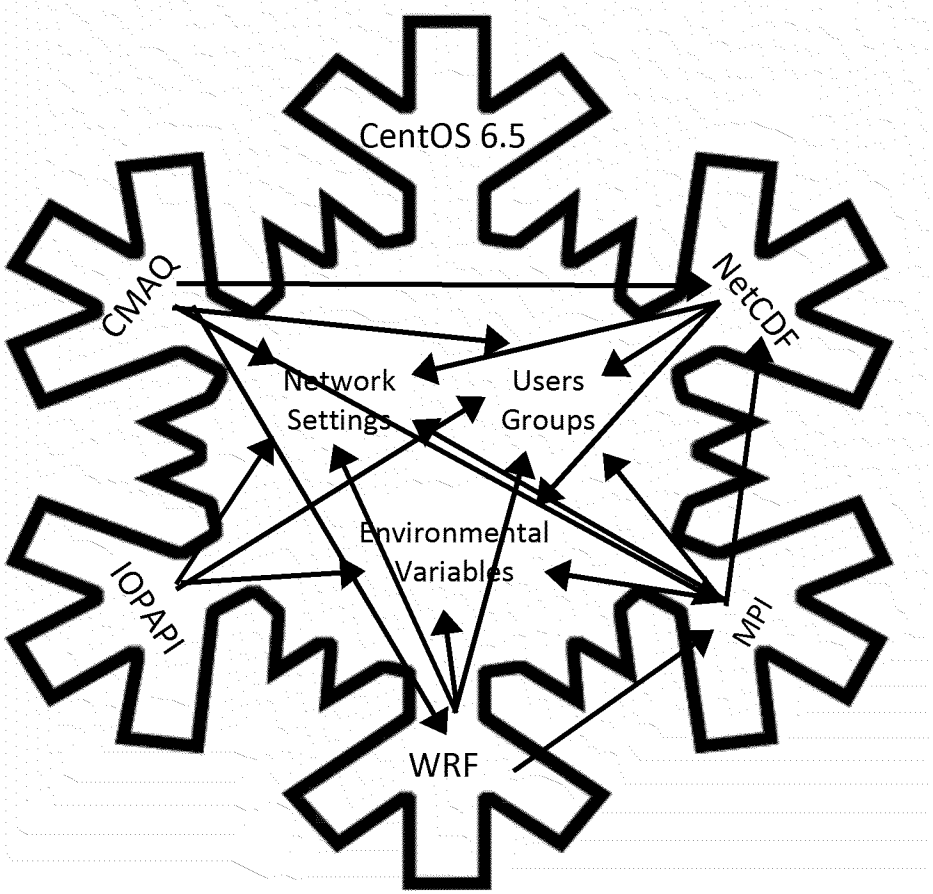
- Scientists spend time debugging compilers, libraries, and run scripts, instead of doing scientific analysis.
- Model set-up that works on one machine may not work on another due to OS and compiler issues.
- *Not scalable* – you are limited by the CPUs and RAM on your (in-house) machine.
- *Not automated* – each preprocessor and model has to be run separately, with potential for errors from inconsistent input specifications.
- Hard to do forecasts or rapid response modeling.
- ***Overall, it takes too long, requires too much labor, and is needlessly complicated.***

Similar Problems in the Software/IT World: The Snowflake Server

We need to update the version of NetCDF

We need to add a new user

HEARTBLEED!?!?!?!?!?! ALSO FLASH!?!?!?!?!

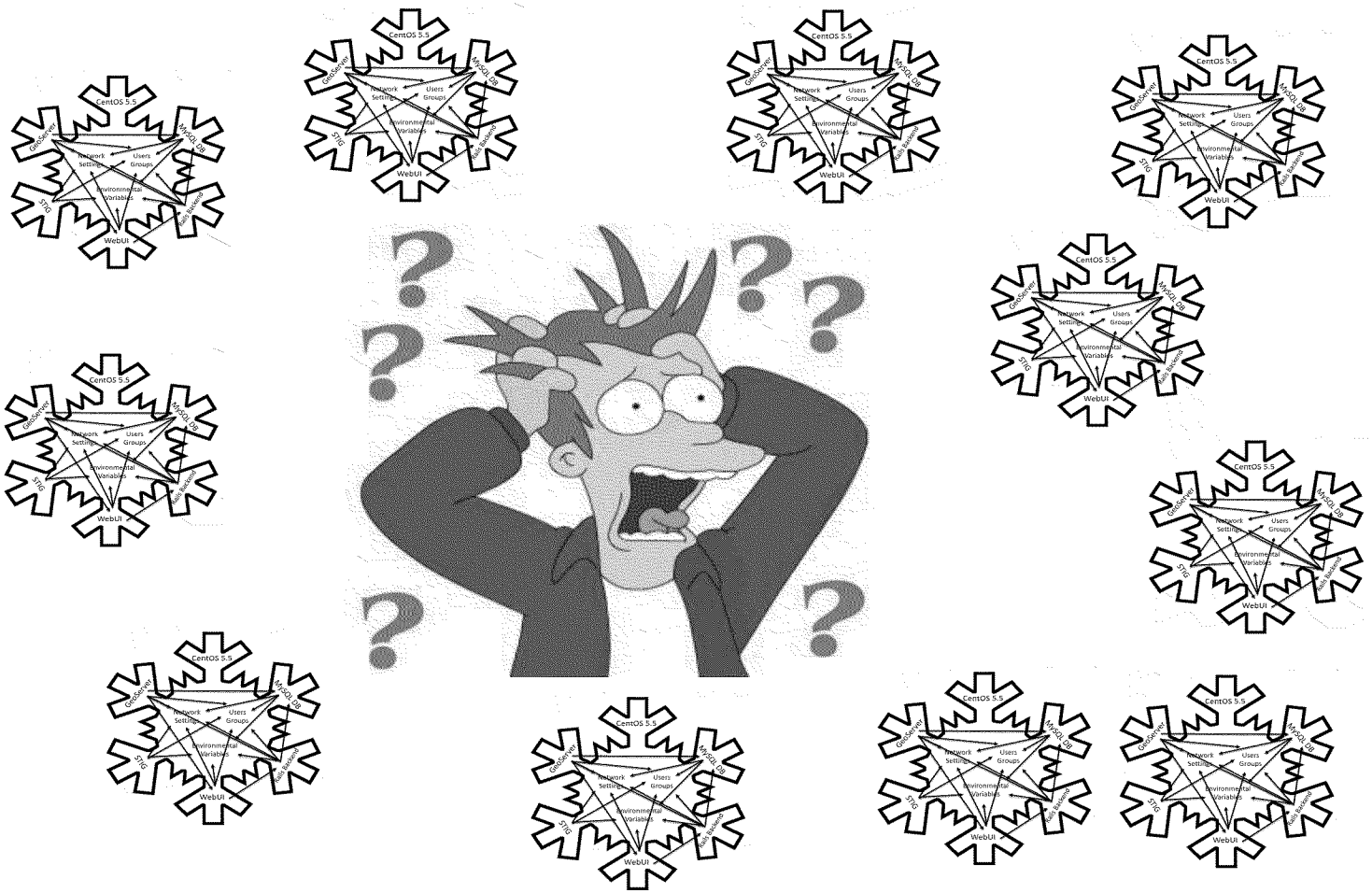


This hard drive sounds like a cement truck today…

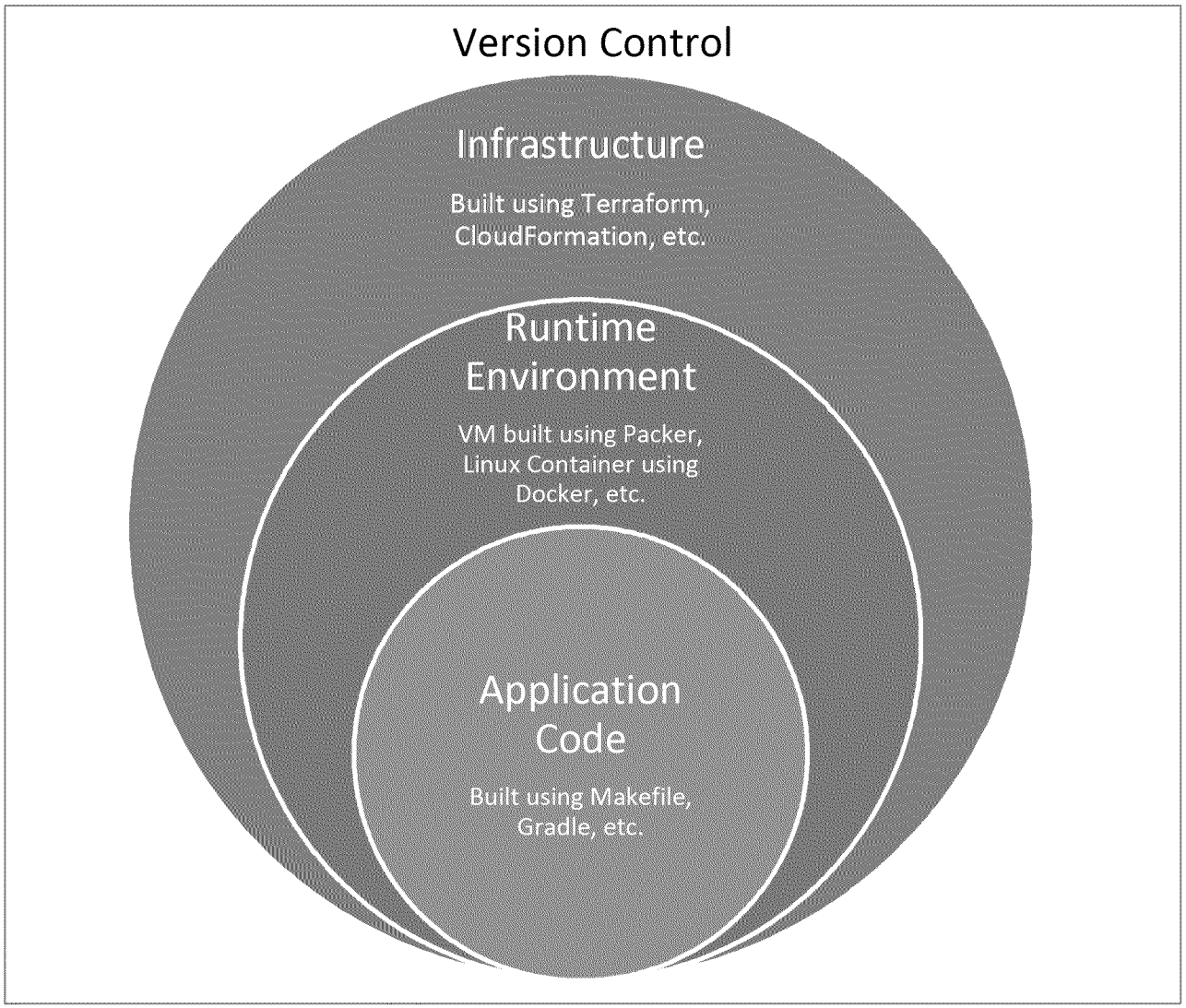
Can we upgrade this to CentOS 7 by tomorrow?

Can we optimize this whole thing?

Cloud gives you many servers, but if they are all unique snowflakes…



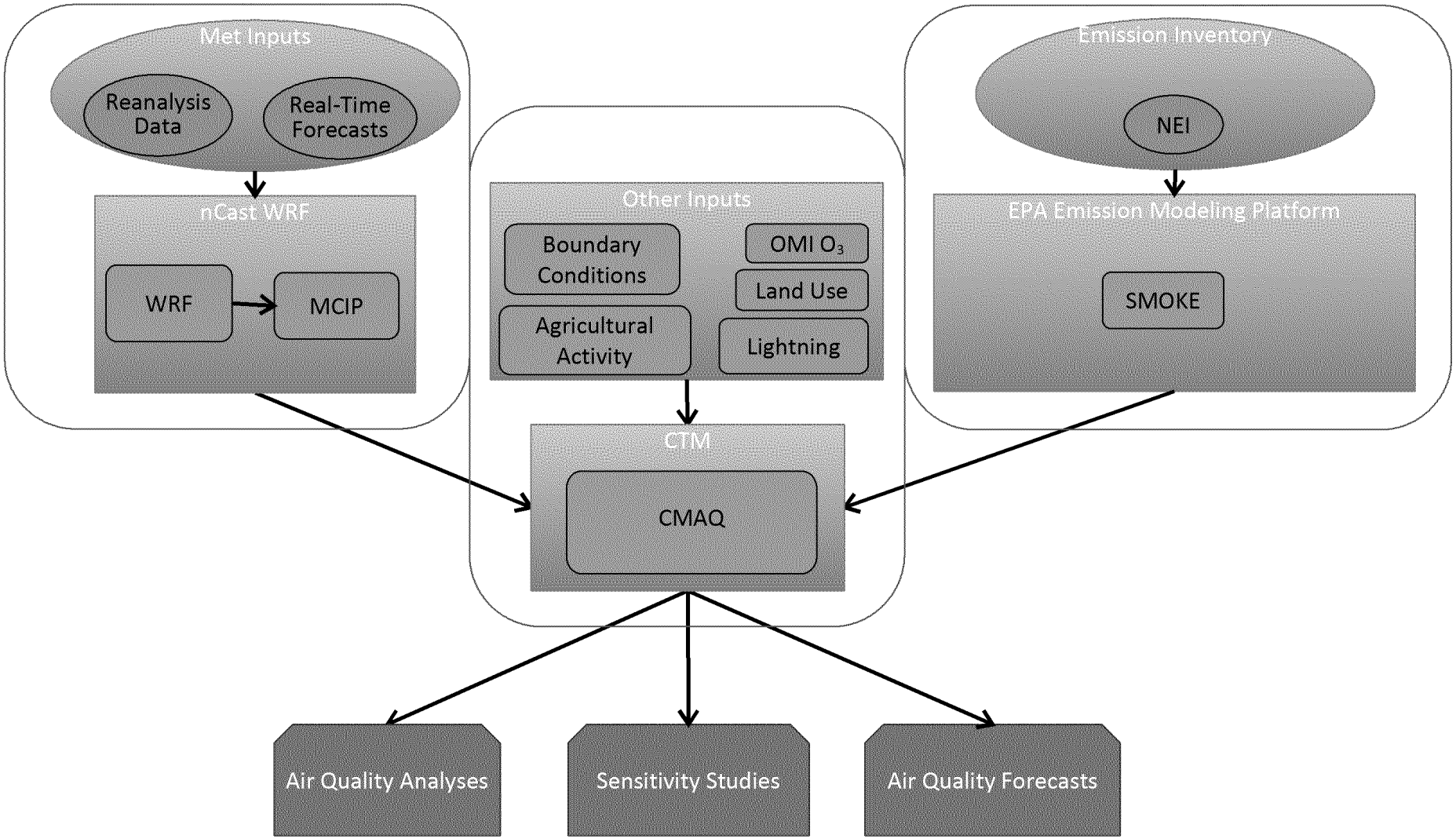
Solution: Infrastructure as Code



Our New Modeling Approach

- ✓ Scalable System on Amazon Web Services (AWS) Cloud
  - Run arbitrary number of model runs on identical machines.
  - Configuration (OS, compilers, etc.) saved as Terraform modules.
  - Only pay for the CPU time you actually use.
- ✓ Automated System
  - Post-processing of WRF with MCIP done automatically.
  - Emissions for different sectors are processed automatically and are merged into model-ready files.
  - Pre-processors for CMAQ in-line emissions and boundary conditions run automatically.
- ✓ Unified System
  - All components run on a common architecture using common XML input specification file.
  - Removes chances for inconsistencies in specifications.

AQcast System for Continental US

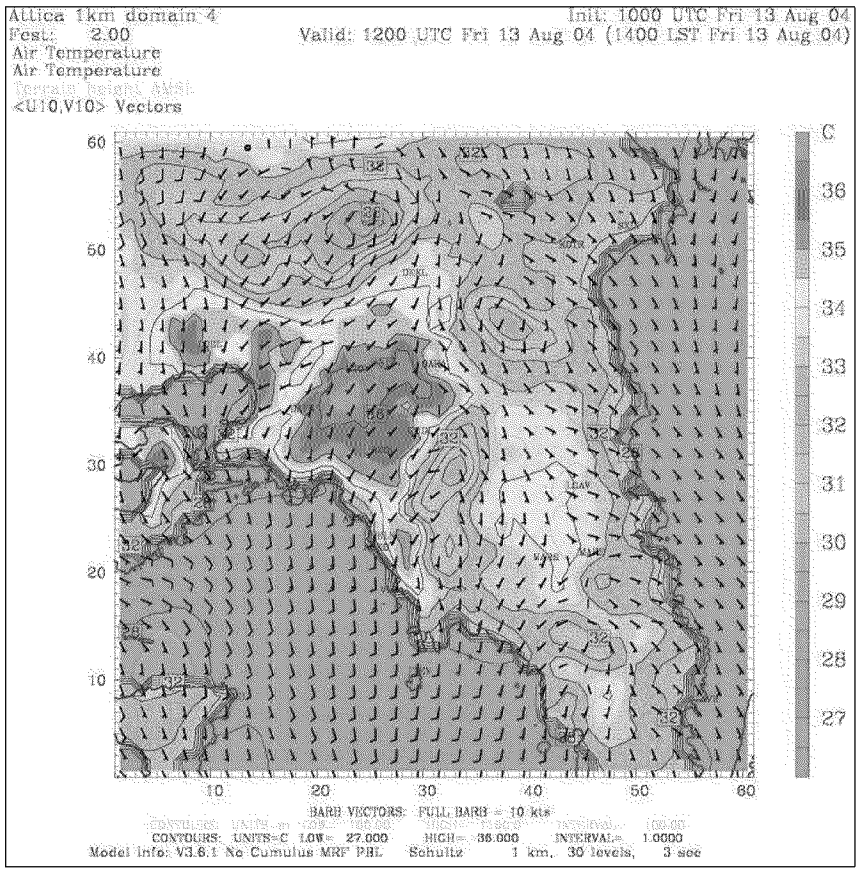


Meteorological Component

- WRF-ARW v3.8
- nCast automatically collects needed input data from reanalyses or forecasts (e.g., NARR, NAM-12, NCEP FNL)
- MCIP integrated into nCast, so that MCIP post-processing is automatic.
- All other components read grid information from Met component.

nCast WRF Forecasts for Greek Olympics

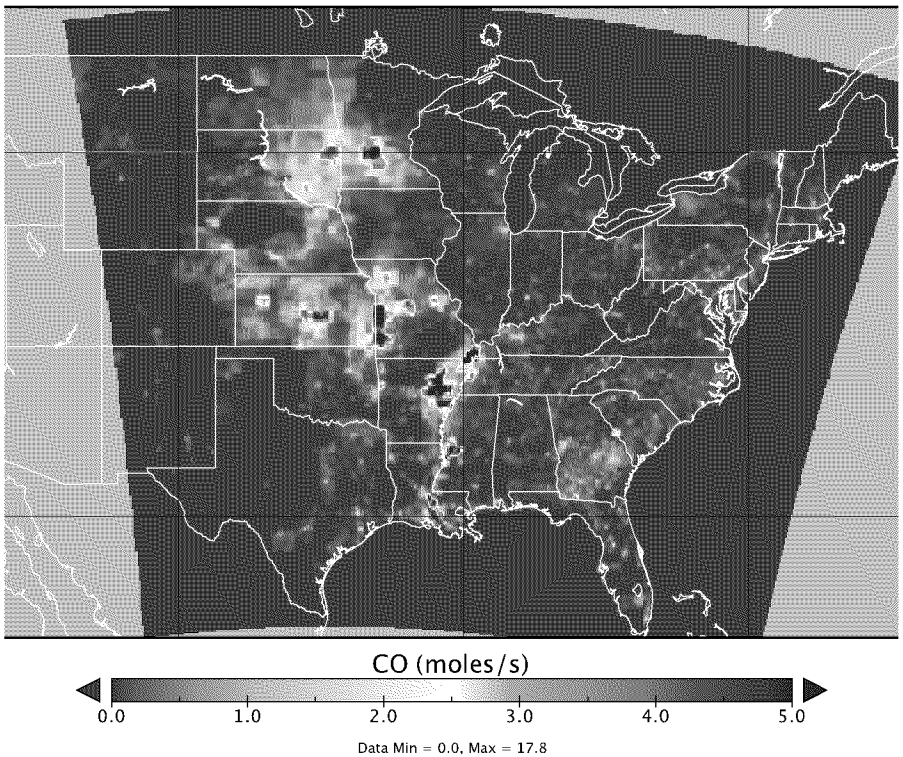
Fig. 5: *Example of nowcasting system graphic showing forecast of 2-m temperature (shaded, degrees C) and 10-m wind field (using standard notation of one barb=5ms^{-1}).*



Emission Component

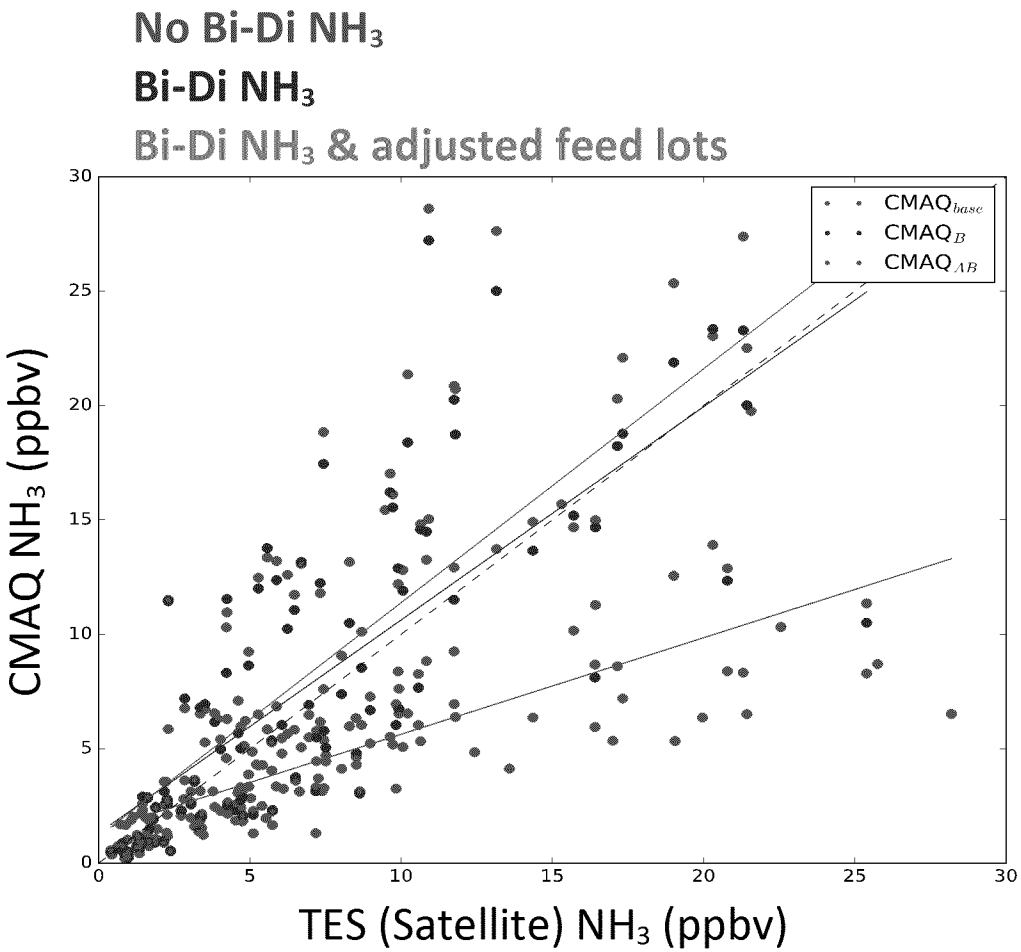
- EPA 2011 NEI Emission Modeling Platform (SMOKE)
- Running of platform scripts debugged, simplified, and automated.
- Can be run for a given subset of days.
- Spatial surrogate generation included.

Non-point source CO, June 3, 2013



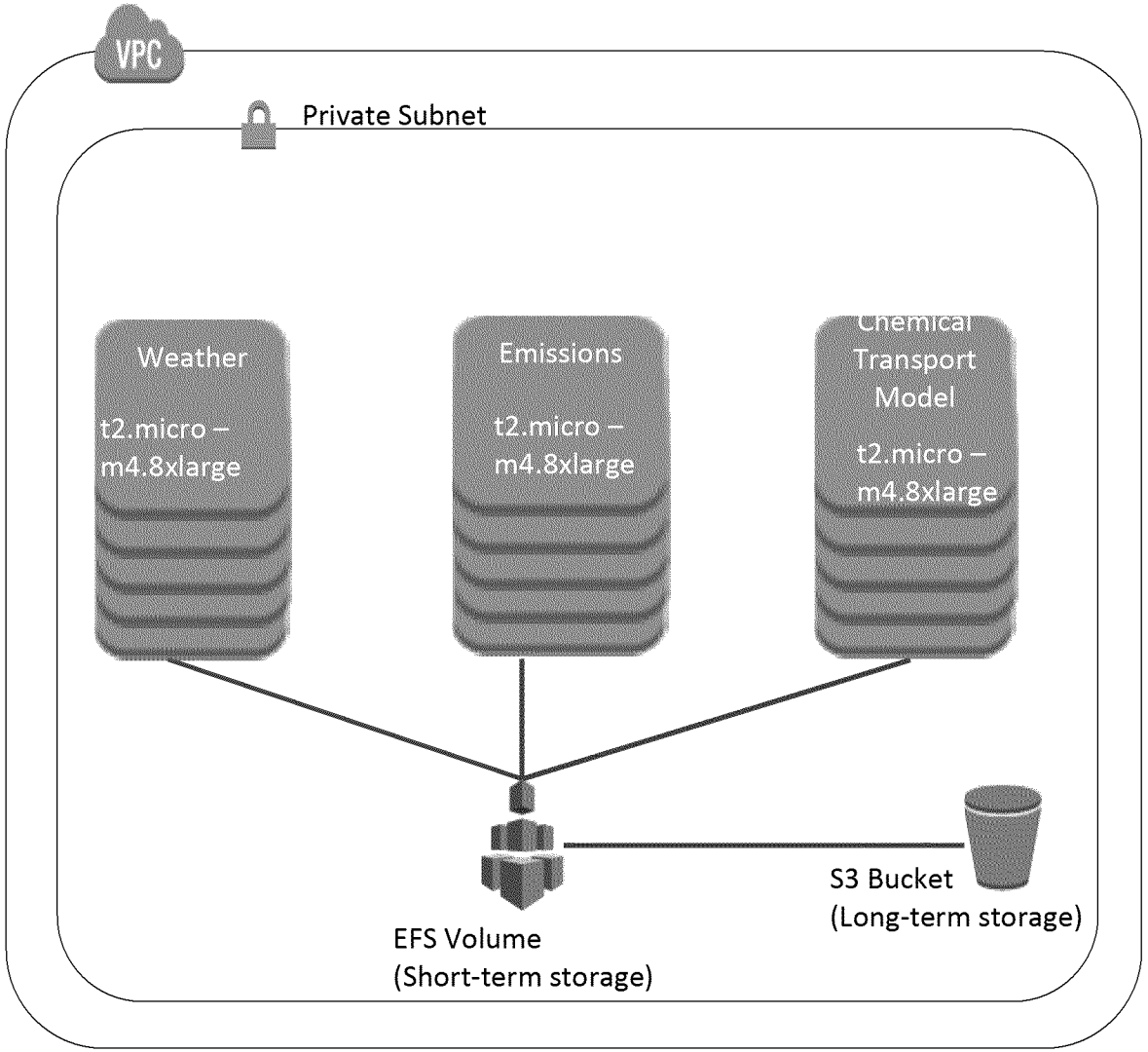
CTM Component

- CMAQ v5.0.1
- Initial and boundary conditions from NCAR MOZART4 forecasts
  - GEOS-Chem converters also available
- Preprocessors automatically make needed inputs for:
  - In-line Dust Emissions
  - In-Line Biogenic Emissions
  - In-Line Lightning NO_x
  - Bi-Directional NH_3



From Lonsdale et al., ACP, 2017.

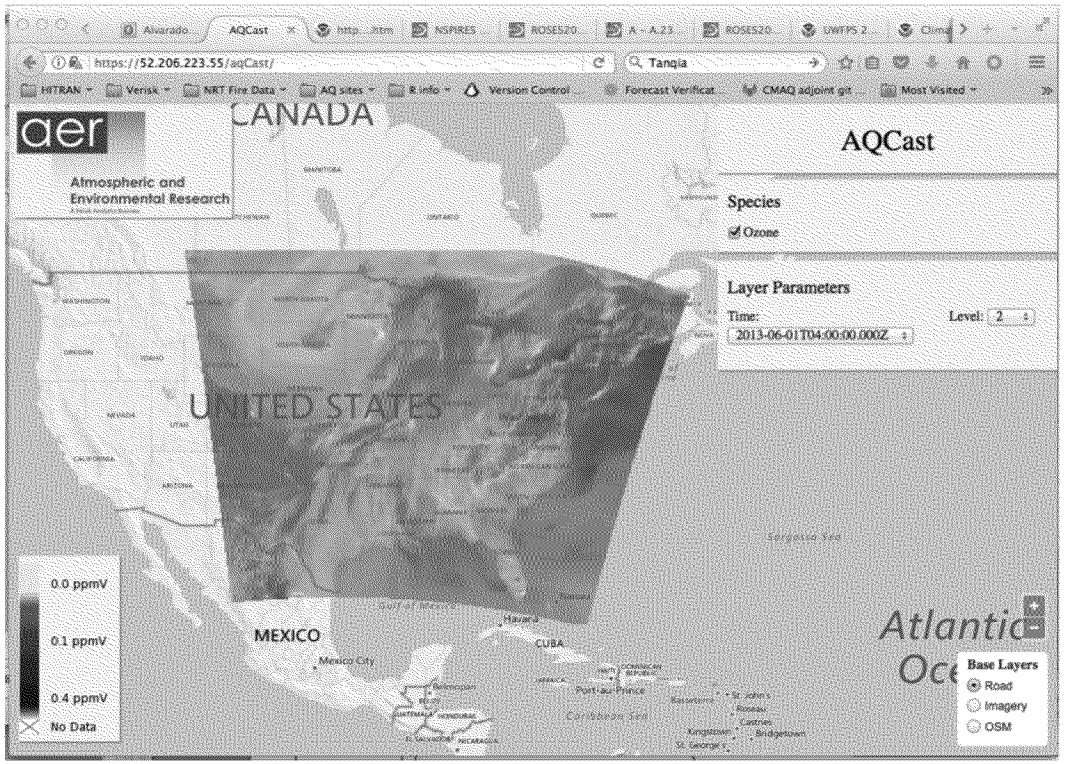
Implementation on Amazon Web Services



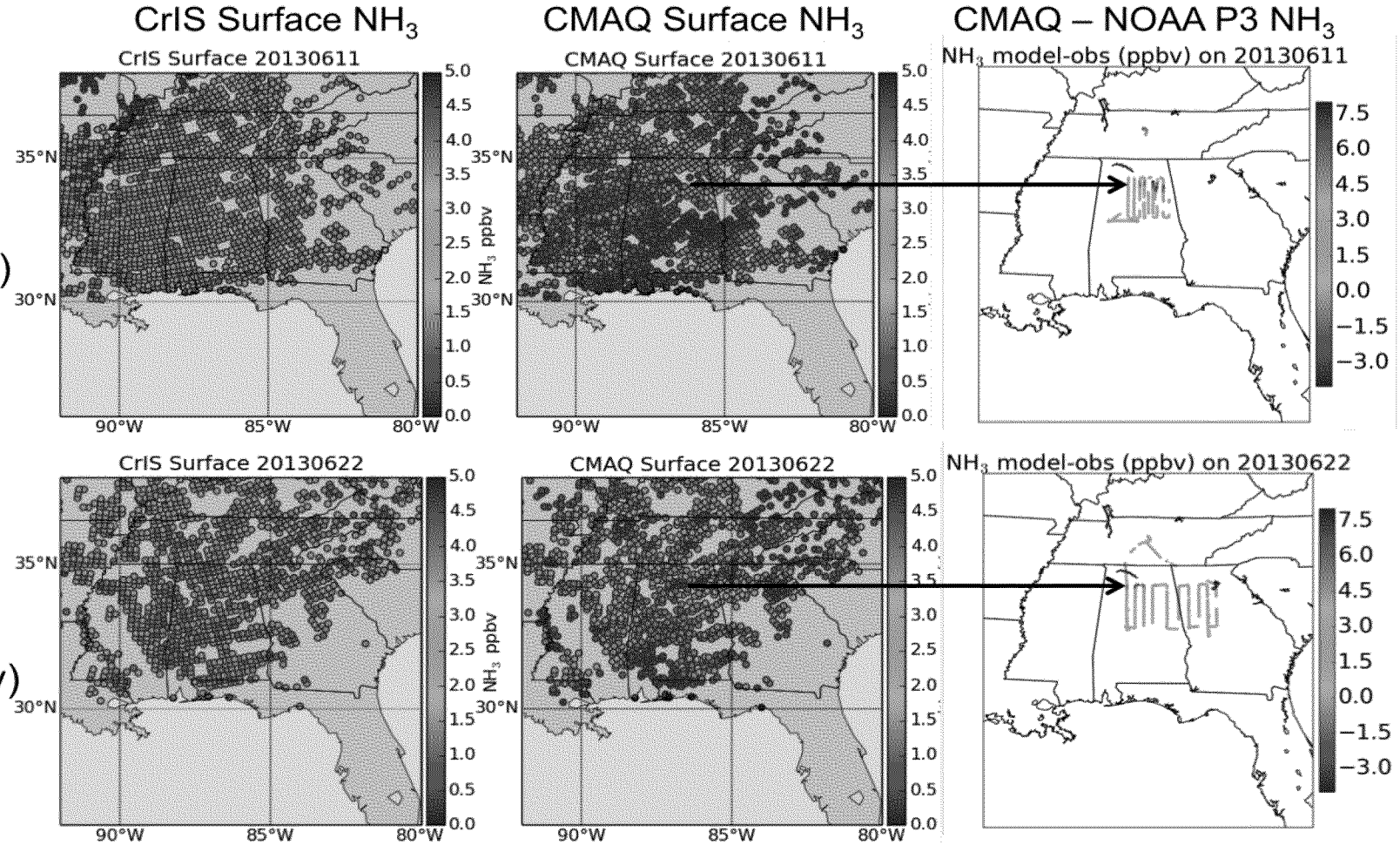
- Components can be run separately or together.
- Preexisting input met or emission files used by uploading them into the EFS volume.
- Currently, starting and stopping machines and the transfer the files between storage volumes is done by the user, but we are working to automate that as well.

QA/QC and Analysis

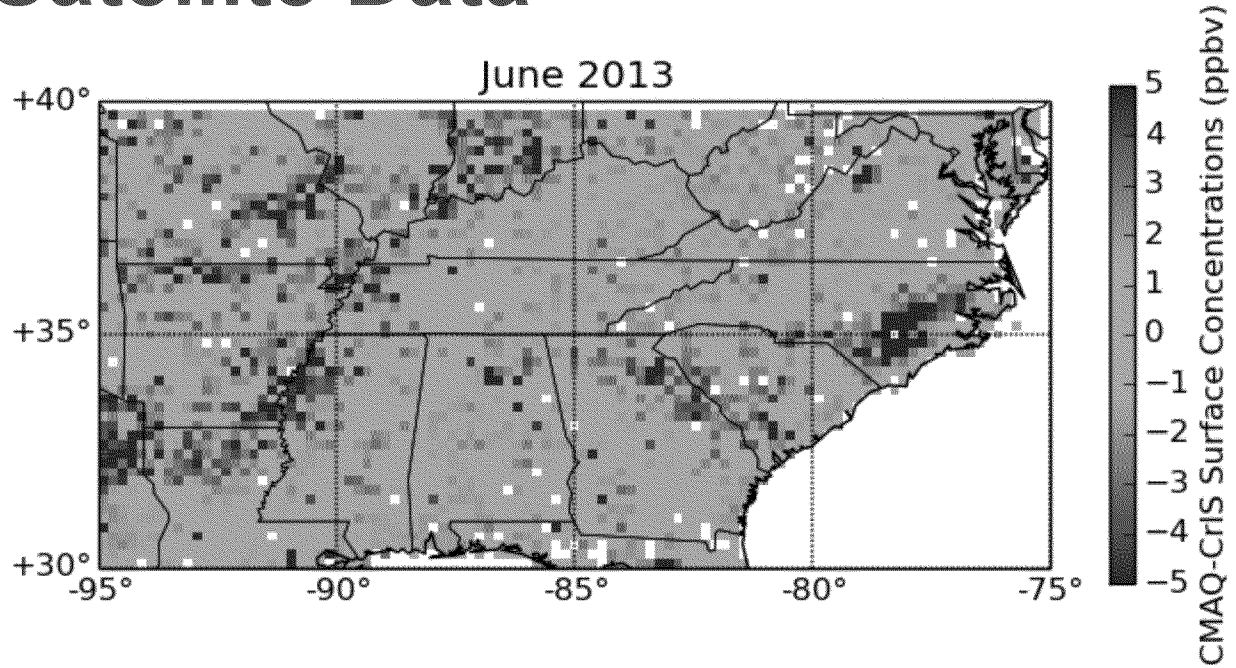
- Data can be moved off the cloud for analysis, but Amazon charges a fee (~$90 per TB).
- QA/QC and analysis scripts can be run on small, one-processor instances on the cloud.
- Using open-source tools (Python, R) avoids license issues.
- Can also set up a server (e.g., THREDDS) to make data accessible via netCDF viewers like Panoply or via the web (see right).



Back to the Science: Evaluating NH_3 Emissions Using Satellite Data



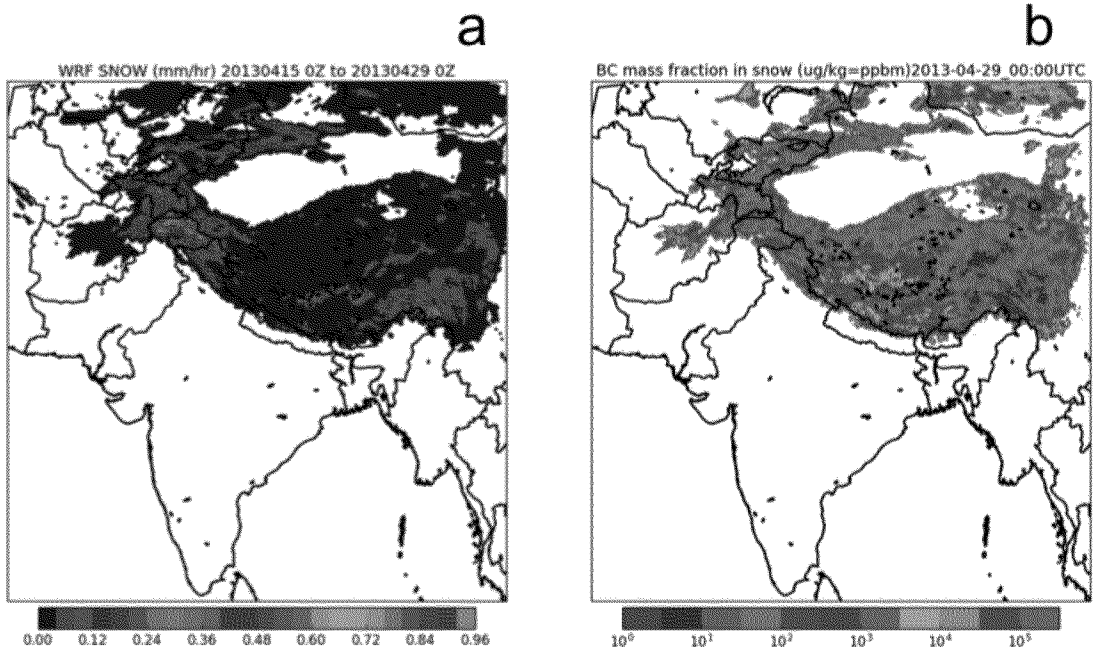
Cloud Extension: *Improving* NH_3 Emissions with Satellite Data



- Using finite-difference, iterative mass balance approach (Lamsal et al., GRL, 2011).
- Requires two model runs per iteration – baseline and perturbed.
- Calculate emission scaling factors and apply to next iteration.
- *Ideally needs two identical clusters running in parallel, taking advantage of cloud computing.*

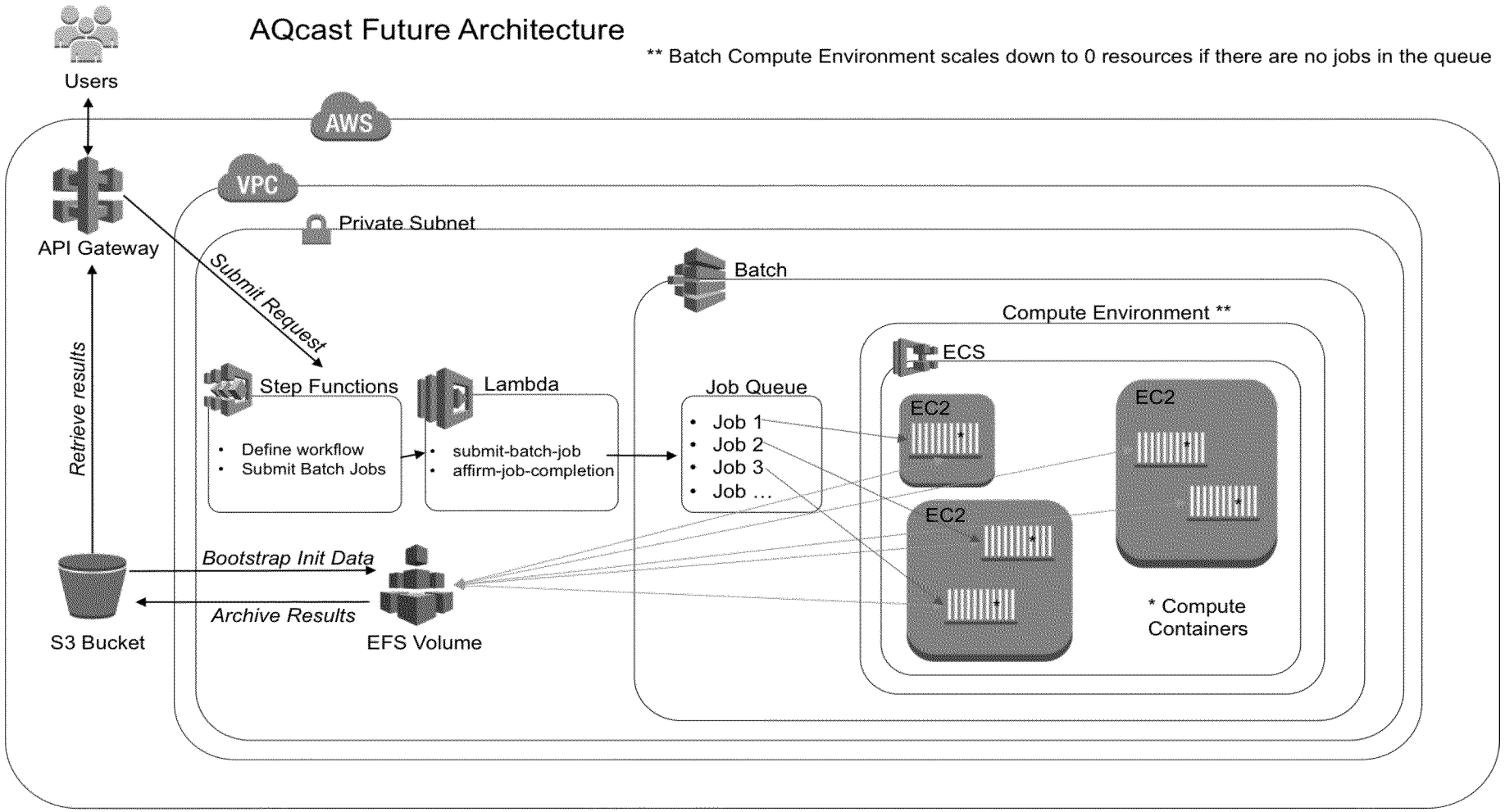
Future Plans: International and Streamlined

- Incorporate CAMx as an alternative CTM.
- Incorporate WRF-Chem for meteorology-pollution interaction studies.
- Expand emissions component to cover the globe.
- Build web-based interface for submitting jobs.



Example from WRF-Chem run on the cloud. Runs were performed for 2 emission scenarios, three phases of ENSO, and four months per year (one per season), for 24 total runs. (a) Average snowfall (mm/hr). (b) Mass ratio of BC deposition to snowfall (µg BC/kg snow).

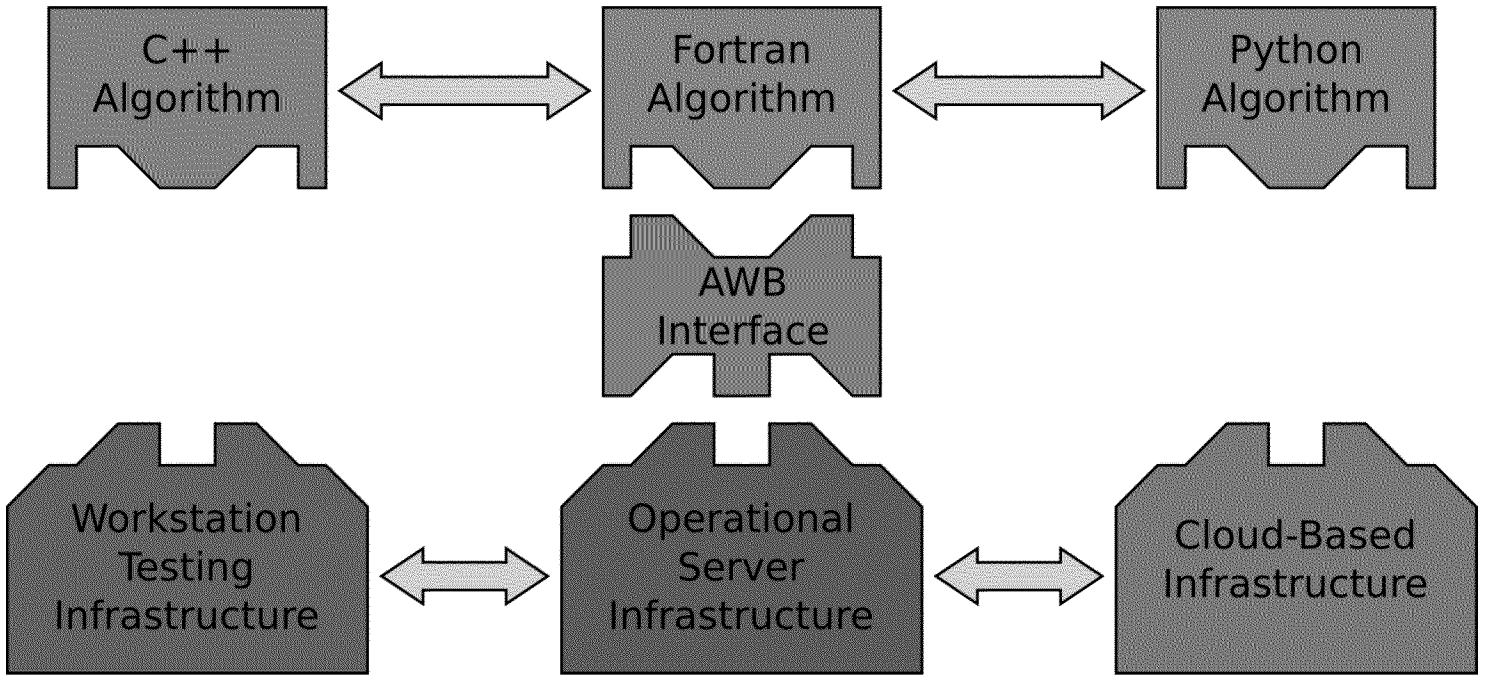
Planned Architecture Improvements



Summary

- Cloud computing allows you to quickly run an arbitrarily large number of AQ model simulations.
- Saving the infrastructure as software allows you to create identical machines and easily maintain them.
- We used these principles to create a ***scalable, automated system*** for air quality modeling.
  - Reduces the time and labor to go from idea to analysis.
  - Reduces learning curve for new users.
- Future plans include allowing jobs to be submitted via the web, adding additional CTMs, and allowing modeling outside of US.
- Questions? Email malvarad@aer.com

Future: Modular Algorithm Interfaces



- Initially designed for satellite data processing
- Standardized interface allows algorithms and infrastructure to be swapped without code changes
- Algorithms in different languages run in the same environment
- Users can change/add algorithms using standard interfaces